**Transport\_event analysis Sebastian Acciarito**

**Data Description**

This data has been sourced from Vicroads Open Data. It contains the traffic volumes for freeways (excluding toll roads) and arterial roads in Victoria. It provides information on the annual average daily traffic volume, including the number of commercial vehicles. Values within the data have been derived from traffic surveys or produced from estimates.

**Data Preprocessing & Cleaning**

It was determined that the identification columns (OBJECTID, LOCATION\_ID, HMGNS\_FLOW\_ID, HMGNS\_LINK\_ID), the label, year and HMGNS\_LINK\_DESC columns did not provide useful information and were removed. Further, ALLVEH\_AMPEAK and ALLVEH\_AMPEAK were removed due to missing variables.

We chose to add a new column known as Street, by striping the street name of each roadwork events from the HMGNS\_LINK\_DESC column before removing it. We chose to add this column to be able to label the points, when visualising the clusters. Lastly we chose to reduce the number of observations to 1000 (500 as a training subset and 500 as a test subset) for two reasons. We needed to reduce the dataset significantly due to the computational limitations of our laptops. Even when enabling the GPU for computation, initial tests on using the full data set took over 30 minutes. Given that we are testing a variety of different cluster sizes, we strove for efficiency in cluster creation. Secondly, we attempted to run the K-means algorithm using data sets of 1000, 3000, 5000, 8000 observations and found that we achieved the same inferences each time. Hence we chose to use the smallest sample size (1000) as the k-means algorithm appears to be stable with a minimum number of observations. The data set was then reduced to a random selection of 20 observations for clear visualization. Please see the appendix for a full variable list.

**Preliminary Analysis**

From Figure 2 we can see that the road containing the highest daily volume of traffic is heading Westbound. Consequently, the second busiest road in terms of daily traffic volume is heading Eastbound. Both of these figures are approaching 120,000 vehicles per day. More so, these two observations have a low growth rate - smaller than 5% - yet they are still positive indicating they are still growing in volume.

Further, Westbound direction of traffic flow also contains the road with the highest growth rate, approximately 60%. Interestingly, this road also has quite a low level of daily traffic volume. The high growth rate may be caused by factors such as urban development or new infrastructure.

Alternatively, the North-Westbound flow of traffic contains the lowest traffic volume growth rate, approximately -25%. This road also appears to have no, or very little, daily traffic volume. It is difficult to tell from this plot, but it appears Westbound roads and South-Eastbound roads dominate in traffic volume.

Figure 3 shows two roads whose traffic volume consists of only trucks. These roads are headed in Southbound and Northbound directions and have negative growth rates of approximately 15%. We can also distinguish that vehicles other than trucks dominate roads in all flow directions. Interestingly, roads with the lowest percentage of trucks also make the highest growth rate roads.

**Principal Components Analysis (PCA)**

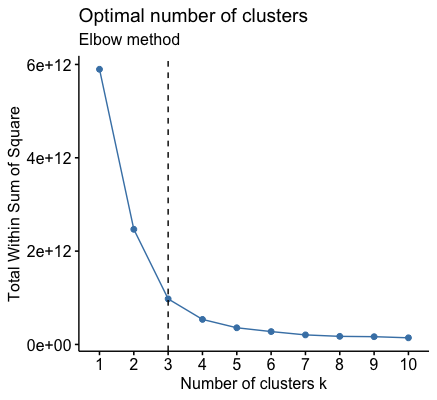
As depicted in figure 4, traffic flow directions that are close to each other, such as North-westbound, South-Westbound and Westbound, can be perceived as similar in terms of their traffic flows and volumes. Meanwhile, North-Eastbound and Southbound traffic flows have no correlation in terms of their traffic flows and volumes. Interestingly, two way yearly volume for all vehicles is strongly positively correlated with Eastbound, Northbound and South-Eastbound roads. On the other hand, the percentage of trucks that make up all vehicle volume is either uncorrelated or negatively correlated with all road flow directions. Figure 5 shows the importance of each principal component calculated. As shown, PC1 explains 99.99% of all the variance between each variable.

**Building our clusters (K-means)**

Prior to building our clusters through K-means, we must determine the number of clusters required for robustness. The following measures of robustness will be used:

Total Within Sum of Squares, Rand index and Adjusted Rand Index.

For the Total Within Sum of Squares method, we compute clusters, and save the Total Within Sum of Squares as separate objects. We then plot the objects on a linear graph as illustrated below. The ‘Elbow method’ as observed in Figure 6, suggests that 3 clusters is optimum.

****

*Figure 6*

The Rand index determines the total agreement between clusters, when two random observations are chosen. That is, the likelihood of choosing two observations in the same cluster.

|  |  |
| --- | --- |
| **Rand Indexes** | **Probability of Agreement** |
| Rand Index for 2 Clusters | 0.9120561 |
| Rand Index for 3 Clusters | 0.5037595 |
| Rand Index for 4 Clusters | 0.5088737 |
| Rand Index for 5 Clusters | 0.5175070 |

*Figure 7*

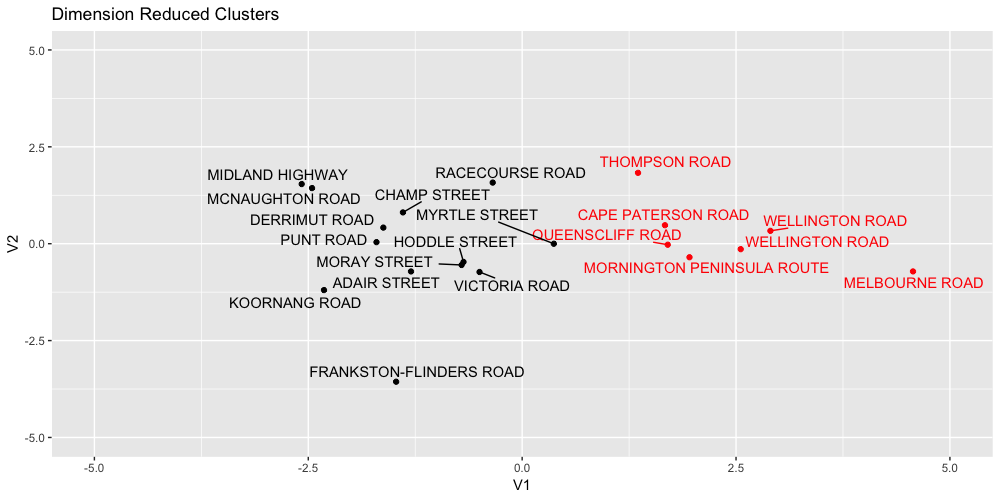
Probabilities in Figure 7 indicate that we are likely to achieve the highest probability of agreement using a 2 cluster solution, with a probability of 91.2056%. This infers that we should use a 2 cluster solution, and is in contrast to the Total Within Sum of Squares method that suggests a 3 cluster solution.

Lastly, we included Adjusted Rand Index to see if we obtain a comparable result to the Rand Index after accounting for agreement between observations being a result of chance. The Adjusted Rand Index (Figure 8) for suggests that a 2 cluster solution appears to have the highest probability of agreement -- 5.39%. However, the low probabilities of the Adjusted Rand Index infer that perhaps chance largely influences how observations are clustered. Nonetheless, we will still infer that a 2 cluster solution is still the optimal number of clusters for a dataset, as both the Rand and Adjusted Rand indexes suggest that the highest probability of agreement can be achieved using a 2 cluster solution.

|  |  |
| --- | --- |
| **Adjusted Rand Indexes** | **Probability of Agreement** |
| Adjusted Rand Index for 2 Clusters | 0.0539181 |
| Adjusted Rand Index for 3 Clusters | 0.0075814 |
| Adjusted Rand Index for 4 Clusters | 0.0108335 |
| Adjusted Rand Index for 5 Clusters | 0.0085158 |

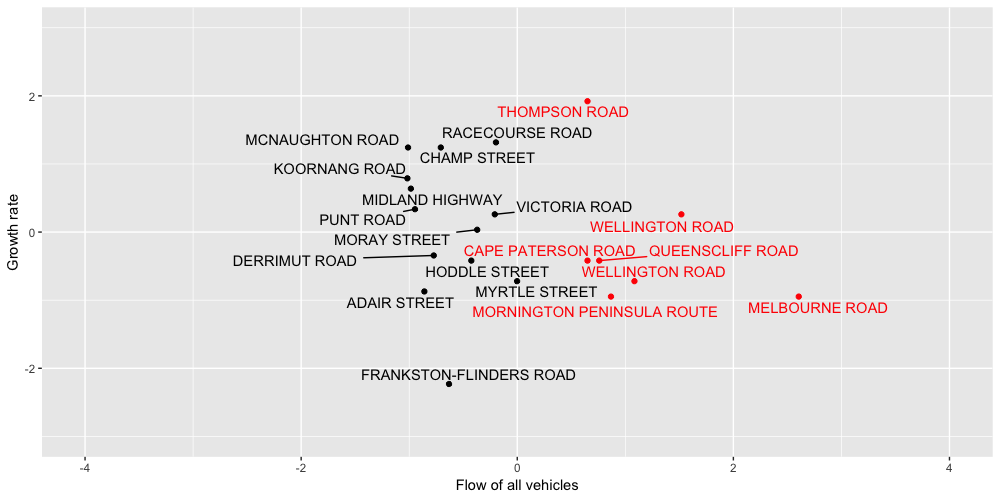
*Figure 8*

We will now analyse the clusters visually to determine if they are associated with different parts of Melbourne**.** We have chosen to use multidimensional scaling as means of summarizing the variables used to produced the cluster into two clear variables(V1,V2). This visualization illustrates only 20 observations, as the full dataset could not be graphed to display both the points and streets clearly. The Multidimensional visualization for the full 500 can be found in the appendix.



*Figure 9*

Figure 9 suggests that there are two clear clusters among the data. In fact we see that there appears to be a significant division of space between the clusters. Through the use of Google maps, we were able to determine the that the first cluster (black points) are considered to be road work events in high density/low traffic areas whilst the second cluster (red points) appear to be road work events in low density/high traffic areas.



*Figure 10*

This cluster classification is exemplified in Figure 10, where we see that black points have a positive growth rate with negative traffic flow. There are exceptions to this rule with Derrimut Road, Adair Street, Frankston-Flinders road indicating that these road types have both negative growth and negative traffic flow. However this still fits the black cluster well as it indicates that there is very little traffic flow due to the high density of these areas. In comparison, the red points appear to have a negative growth rate but positive traffic flow. However it is important to note that Thompson Road and Wellington Road appear to have positive growth and positive traffic flow. The reason for this is that both Wellington and Thompson road appears to be a major roads connecting high density residential areas.

Dendrogram analysis allows us to visualise clusters in terms of growth rate and business.

We have used this tool to depict the quietest 20 streets (figure 12), the busiest 20 streets (figure 13), 20 streets with the lowest growth (figure 14), and 20 streets with the highest growth.

Unsurprisingly, the Monash and Westgate freeways made up the majority of the busiest streets.

**Validating our clusters (MANOVA)**

Before using MANOVA to test the significance of our clusters, we must first test if the following three assumptions hold:

1. Observations are independently sampled:
   1. As we did not collect the data, we can only assume that this assumption holds
2. The variance covariance matrix is the same for all groups:

H0: variance covariance matrices are the same across all groups

H1: at least one variance covariance matrix is different

alpha = 0.01 Reject H0 if p-value < alpha

|  |  |  |
| --- | --- | --- |
| Chi-Square (approx.) | Degrees of Freedom | P-value |
| 619.19 | 28 | 2.2e-16 |

Since the p-value (2.2e-16) is < alpha (0.01), we reject H0 and conclude there is sufficient evidence to suggest variances are not homogenous between groups at the 1% level of significance. As we have determined variances between groups are not homogeneous, Pillai’s trace will be used to come to a conclusion due to its robustness.

1. The data have a multivariate normal distribution:

|  |  |  |  |
| --- | --- | --- | --- |
| Test | Statistic | P-Value | Result |
| Mardia Skewness | 4313.2104 | 0.0000 | NO |
| Mardia Kurtosis | 111.6846 | 0.0000 | NO |
| MVN | NA | NA | NO |

H0: data is multivariate normal H1: data is not multivariate normal

alpha = 0.01 Reject H0 if p-value < alpha

Since the p-value (0.0000) is < alpha (0.01), we reject H0 and conclude there is sufficient evidence to suggest the data does not have a multivariate normal distribution. This implies that we are unable to rely on any asymptotic arguments made, and can only use MANOVA as a diagnostic tool.

Now that we have determined that variance covariance matrix is not identical across groups and that the data is not multivariate normal, we will use Pillai’s trace to test if he means of clusters 1 and 2 are statistically different.

\*When this is put into a word doc I’ll fix the presentation of the tests below\*

H0: Mu\_Cluster1 = Mu\_Cluster2 H1: The mean of at least one cluster is not the same as another

alpha = 0.01 Reject H0 if p-value < alpha

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | DF | Pillai | F-Statistic | DF | Den. DF | PR(>F) |
| X | 1 | 0.69402 | 159.42 | 7 | 492 | 2.2e-16 |
|  |  | Roy |  |  |  |  |
| X | 1 | 2.2682 | 159.42 | 7 | 492 | 2.2e-16 |
|  |  | Wilks |  |  |  |  |
| X | 1 | 0.30598 | 159.42 | 7 | 492 | 2.2e-16 |
| Residuals | 498 |  |  |  |  |  |

Since the p-value (2.2e-16) is < alpha (0.01), we reject H0. Under all tests there is sufficient evidence to conclude that the differences between clusters 1 and 2 is significant at the 1% level of significance.

Hence the choice of two clusters appears to be appropriate.

**Limitations:**

The first limitation we encountered was the necessity to reduce our dataset. This reduced the accuracy of our results and made our dataset less representative of the true population. We believe that clustering could have been improved if the ALLVEH\_AMPEAK and ALLVEH\_AMPEAK variables contained all observations, and the repeated observations for specific roads were replaced with a single value per location.

Given that we chose to use a random subset of the data, our clusters may be susceptible to outliers, which could affect the determination of centroids and subsequently how data has been classified.

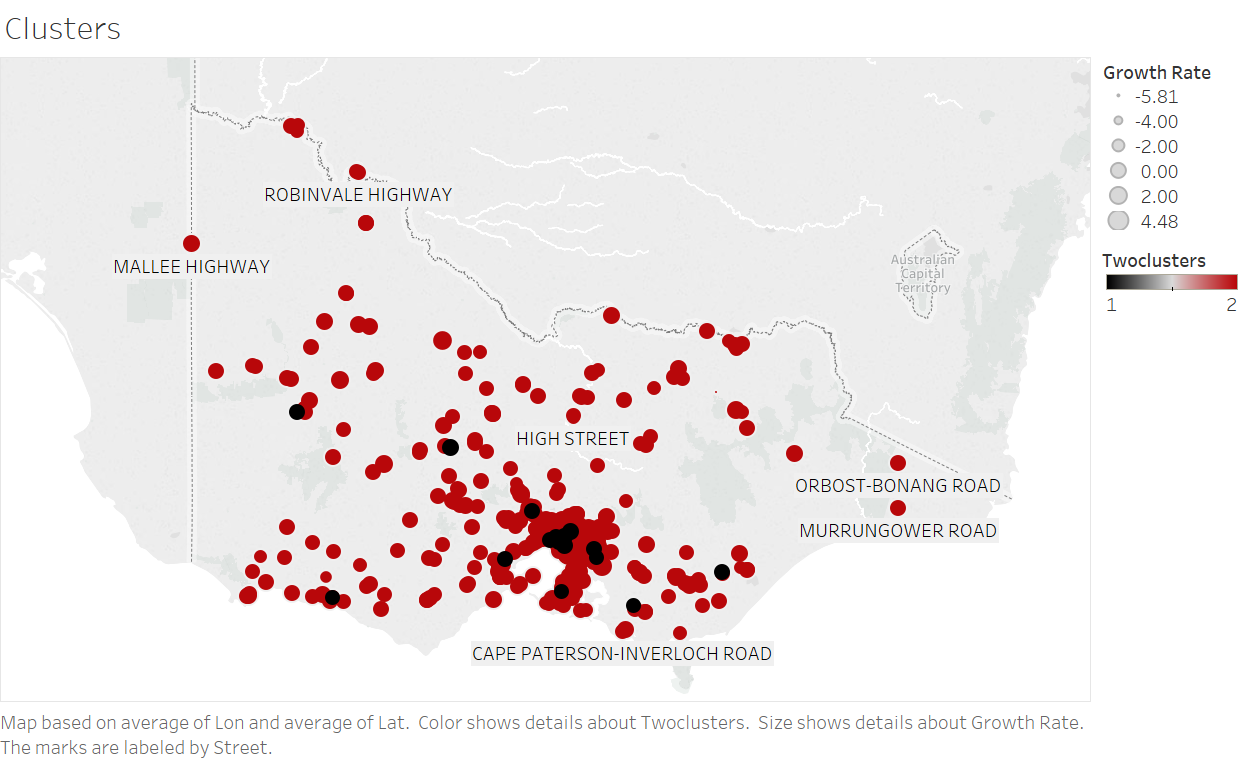
Finally, conflicting results in our initial robustness checks (which fluctuated between 2 and 3 clusters) created overarching uncertainty on our following testing and analysis. However we suspect that if we used more data to build the clusters the Total Within Sum of Squares, Rand Index and Adjusted Rand Index would infer the same clustering solution.

However the most significant limitation of the our analysis is that there is evidence to suggest that clustering has only occurred by chance. Insofar as after using the Adjusted Rand Index, the probability of agreement dropped significantly for all clustering solutions. Thus while we were able to draw insights from this dataset, the same classification may not exist for new data.

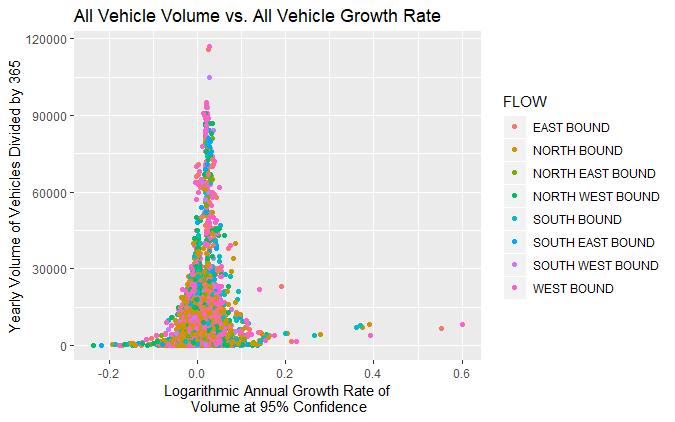
**Appendix:**

* MIDPNT\_LAT: Latitude of midpoint of length segment;
* MIDPNT\_LON: Longitude of midpoint of length segment;
* FLOW: Direction of traffic flow - West Bound, North Bound, South Bound, South East Bound, South West Bound;
* ALLVEHS\_MMW: 24 hour median midweek volume for all vehicles over tuesday, wednesday and thursday;
* ALLVEH\_CALC: Calculation of type used for all vehicle volume. “A” actual volume, “E” estimated volume;
* ALLVEHS\_AADT: Yearly volume for all vehicles divided by 365;
* TRUCKS\_AADT: Yearly volume for trucks divided by 365;
* TRUCK\_CALC: Calculation type used for truck volume. “A” actual volume; “E” estimated volume;
* PER\_TRUCKS: Percentage of trucks that make up the all vehicle volume;
* TWO\_WAY\_AADT: Two way yearly volume for all vehicles divided by 365;
* GROWTH\_RATE: Logarithmic annual growth rate of volume at 95% confidence level. E.g. 1.9% indicate that the homogeneous volume is growing by 1.9% a year;
* CI: Confidence interval on the growth rate. E.g. 0.27% means that the volume on this homogeneous flow is growing at the growth rate +/- 0.27%;
* Streets: Name of streets pertaining to data collected.
* Twoclusters: 2 cluster classification for each observation (levels=1,2)
* Threeclusters: 3 cluster classification for each observation (levels=1,2,3)
* Fourclusters: 4 cluster classification for each observation (levels=1,2,3,4)
* Fiveclusters: 5 cluster classification for each observation (levels=1,2,3,4,5)

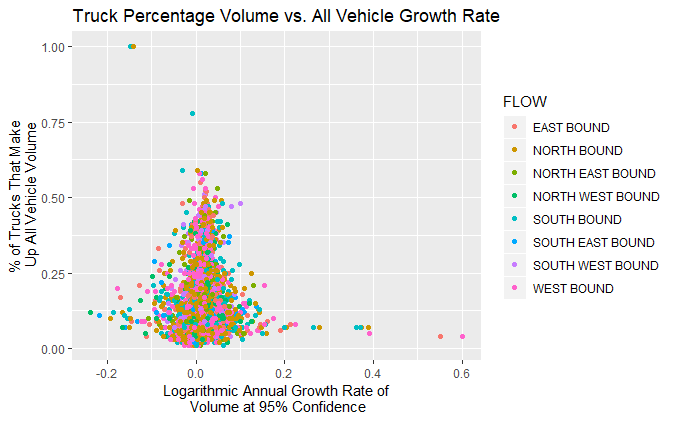
*Figure 1- Victorian road growth rate*



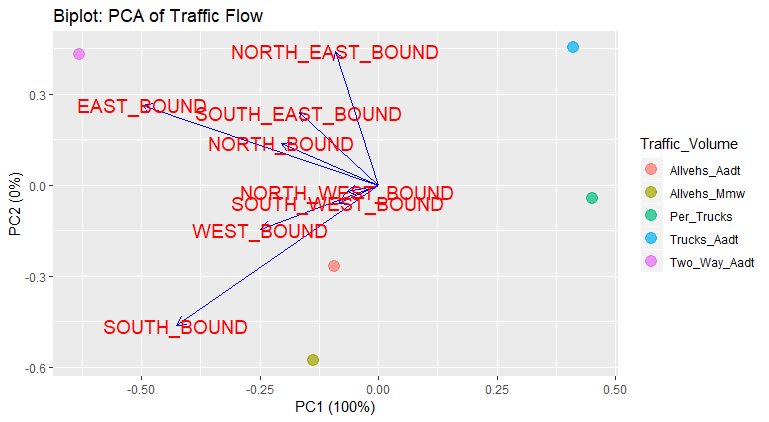
*Figure 2*



*Figure 3*



*Figure 4*

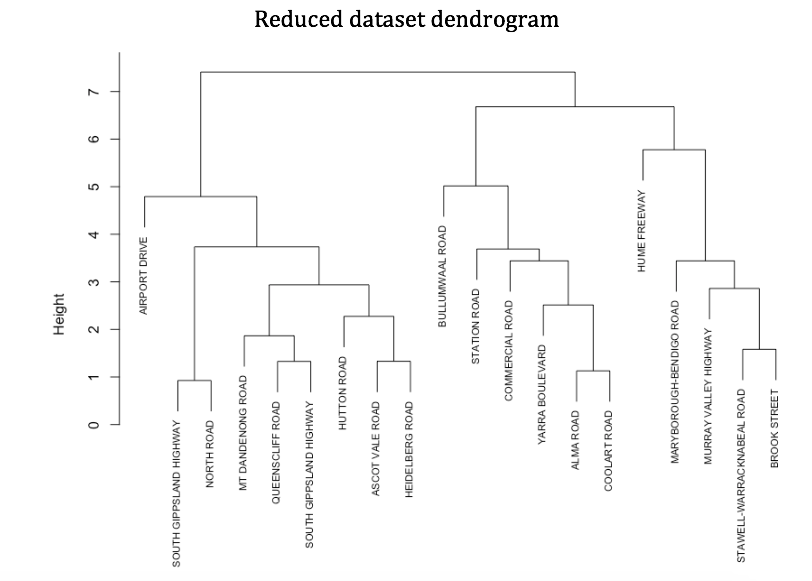


*Figure 5- Importance of components*

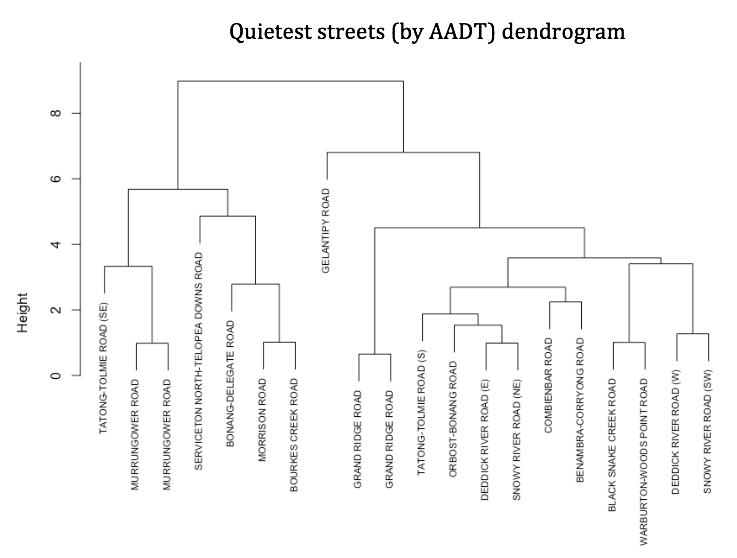
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Component Importance** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** |
| **Standard Deviation** | 2.8283 | 0.02462 | 0.009706 | 0.005595 | ~0 |
| **Proportion of Variance** | 0.9999 | 0.00008 | ~0 | ~0 | ~0 |
| **Cumulative Proportion** | 0.9999 | 0.99998 | 1.0000 | 1.0000 | 1.0000 |

**Figure 5**

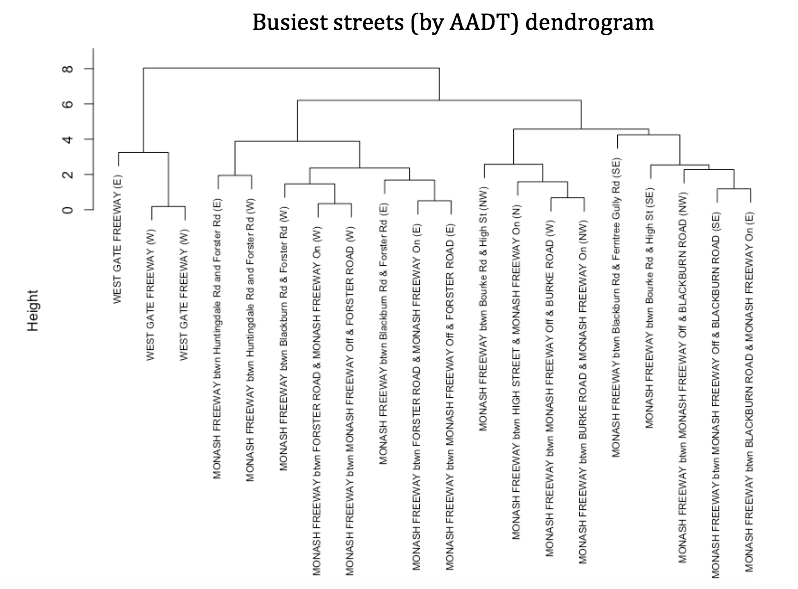
*Figure 11*



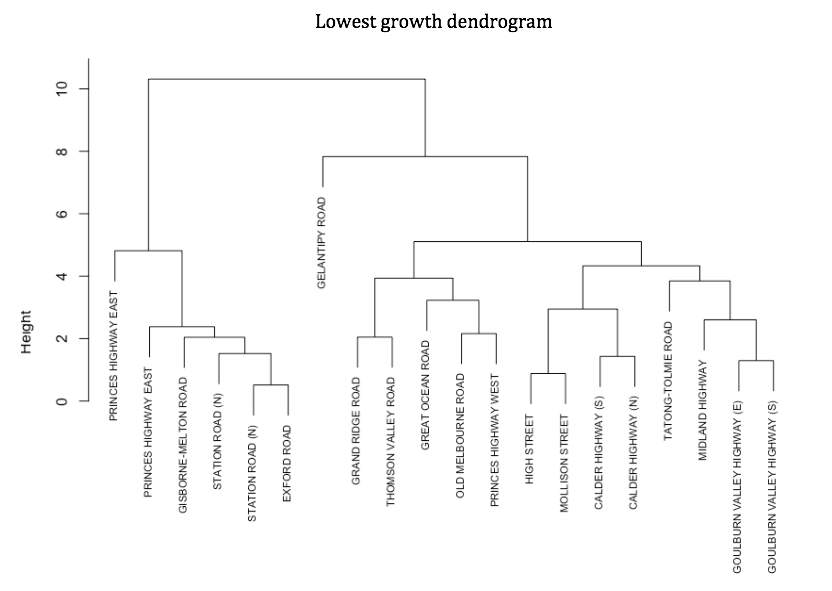
*Figure 12*



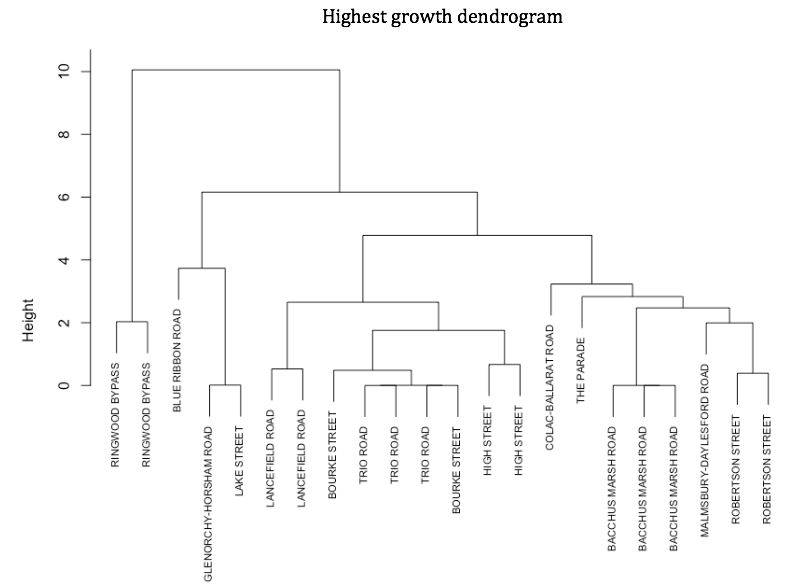
*Figure 13*

****

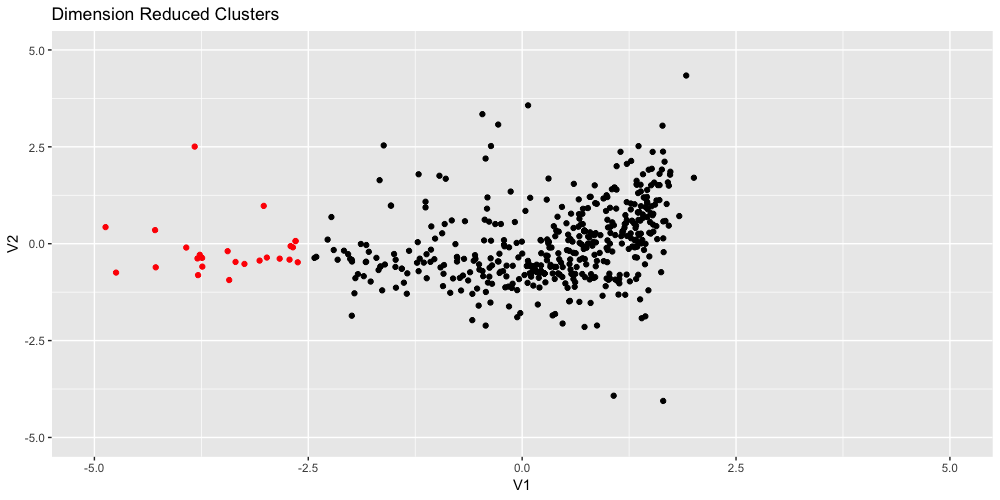
*Figure 14*



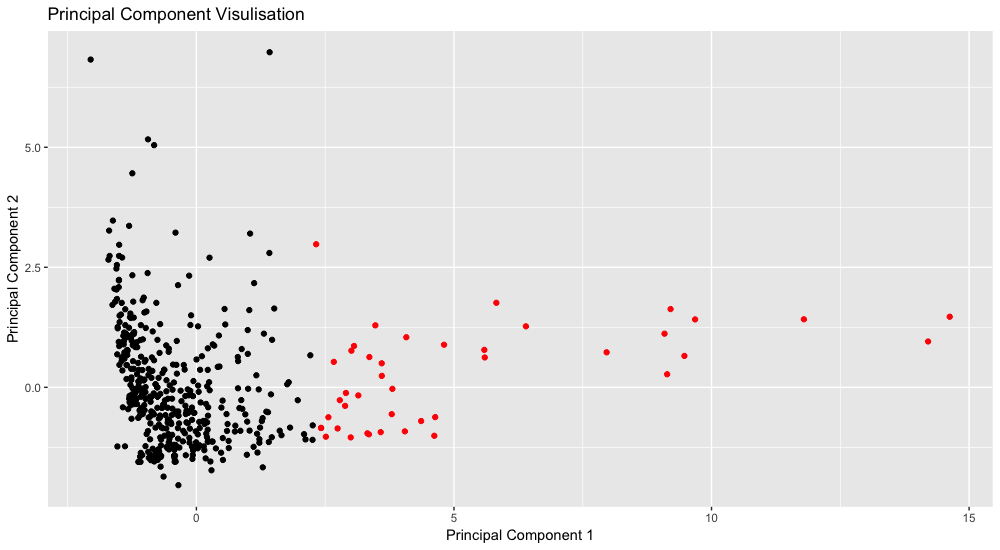
*Figure 15*



*Figure 16*



*Figure 17*



Code for PCA

# PCA

PCAAssign2 %>%

 select\_if(is.numeric) %>%

 prcomp(scale.=TRUE) -> pcaout

summary(pcaout)

biplot(pcaout)

autoplot(prcomp(PCAAssign2[,2:9]), data = PCAAssign2, colour = "Traffic\_Volume",

        loadings = TRUE, loadings.colour = "blue",

        loadings.label = TRUE, loadings.label.size = 5, size = 4, alpha = 0.7) +

 ggtitle('Biplot: PCA of Traffic Flow')

Code for Vehicle Volume Against Vehicle Growth Rate Plot

ggplot(Traffic\_Volume, aes(Traffic\_Volume$GROWTH\_RATE, Traffic\_Volume$ALLVEHS\_AADT)) +

 geom\_point(aes(color = FLOW)) +

 xlab('Logarithmic Annual Growth Rate of\n Volume at 95% Confidence') +

 ylab('Yearly Volume of Vehicles Divided by 365') +

 ggtitle('All Vehicle Volume vs. All Vehicle Growth Rate')

**R code**

**#set working directory**

setwd("~/Documents/UNI/uni 2018 semester 2/High dim data analysis/A2")

**#setting libraries**

library(dplyr)

library(ggplot2)

**#loading data**

tv=read.csv('Traffic\_Volume.csv',header=TRUE,sep=",")

**#Determining road types**

rdt=tv$HMGNS\_LNK\_DESC

Streets=gsub("btwn\\s.\*","",rdt)

tv\_final=cbind(tv,Streets)

**#creating new dataset**

tv\_final=tv\_final[,-c(20:21)]

tv\_final=tv\_final[,-c(16:17)]

tv\_final=tv\_final[,-c(1:2)]

tv\_final=tv\_final[,-c(3:4)]

tv\_final=tv\_final[,-3]

tv\_final=na.omit(tv\_final)

tv\_final=tv\_final[!(is.na(tv\_final$Streets) | tv\_final$Streets==""), ]

**#checking what road types are left**

colnames(tv\_final)

**#reducing rows to 1000 due to computation limit**

tv\_final=tv\_final[1:1000,]

**#randomising**

tv\_final=tv\_final[sample(nrow(tv\_final)),]

#writing csv

write.csv(tv\_final, file="Transport\_Final.csv")

**#Defining Clusters and basic visualization**

setwd("~/Documents/UNI/uni 2018 semester 2/High dim data analysis/A2")

df=read.csv('Transport\_Final.csv')

library(clue)

library(magrittr)

library(dplyr)

library(ggplot2)

library(mclust)

library(knitr)

library(fossil)

library(ggfortify)

library(ggrepel)

df=df[sample(nrow(df)),]

**#removing first strange column**

df=df[,-1]

**#removing categories**

df=df[,-5]

df=df[,-7]

**#splitting out sample and test**

training=df[1:500,]

test=df[501:1000,]

flow\_id=training$FLOW

lat=training$MIDPNT\_LAT

lon=training$MIDPNT\_LON

Street=training$Streets

training=training[,-11]

test=test[,-11]

training=training[,-c(1:3)]

test=test[,-c(1:3)]

**#scaling**

training%>%

 mutate\_if(is.numeric,scale)->training

test%>%

 mutate\_if(is.numeric,scale)->test

sum(is.na(training))

sum(is.na(test))

training=training[sample(nrow(training)),]

test=test[sample(nrow(test)),]

**#Creating different cluster solutions**

c2=kmeans(training,2,nstart=5)

c2m=as.factor(c2$cluster)

c3=kmeans(training,3,nstart=5)

c3m=as.factor(c3$cluster)

c4=kmeans(training,4,nstart=5)

c4m=as.factor(c4$cluster)

c5=kmeans(training,5,nstart=5)

c5m=as.factor(c5$cluster)

**#MDS (Reducing data down to two columns)**

training%>%

dist->delta

mdsout<-cmdscale(delta,eig=TRUE)

mdsout$GOF

mdsout2=cmdscale(delta)

**#PCA**

training%>%

 prcomp->pcaout

principles=as.data.frame(pcaout$x)

pc1=principles$PC1

pc2=principles$PC2

**#saving out dataset for plotting and MANOVA**

training=cbind(training,flow\_id)

training=cbind(training,Street)

training=cbind(training,lat)

training=cbind(training,lon)

training$twoclusters=c2m

training$threeclusters=c3m

training$fourclusters=c4m

training$fiveclusters=c5m

training$pc1=pc1

training$pc2=pc2

write.csv(training,"Training\_data\_with\_clusters.csv")

**#Plotting MDS vectors**

mdsout2%>%

 as\_data\_frame()%>%

 ggplot(aes(x=V1,y=V2))+

 geom\_point(color=c2m)+

 geom\_text\_repel(label=training$Street,color=c2m)+

 theme(legend.position="below")+

 ggtitle('Dimension Reduced Clusters')+

 xlim(c(-5,5))+

 ylim(c(-5,5))

training$flow\_id=as.numeric(training$flow\_id)

**#TESTING 2 clusters**

predicted\_cluster\_memb2<-cl\_predict(c2,test)

cluster\_memb2<-c2$cluster

table(cluster\_memb2)

predicted\_cluster\_memb2

tab2<-table(predicted\_cluster\_memb2,cluster\_memb2)

ari2=adj.rand.index(predicted\_cluster\_memb2,cluster\_memb2)

ri2=rand.index(predicted\_cluster\_memb2,cluster\_memb2)

tab2

**#TESTING 3 clusters**

predicted\_cluster\_memb3<-cl\_predict(c3,test)

cluster\_memb3<-c3$cluster

table(cluster\_memb3)

predicted\_cluster\_memb3

tab3<-table(predicted\_cluster\_memb3,cluster\_memb3)

ari3=adj.rand.index(predicted\_cluster\_memb3,cluster\_memb3)

ri3=rand.index(predicted\_cluster\_memb3,cluster\_memb3)

tab3

**#TESTING 4 Clusters**

predicted\_cluster\_memb4<-cl\_predict(c4,test)

cluster\_memb4<-c4$cluster

table(cluster\_memb4)

predicted\_cluster\_memb4

tab4<-table(predicted\_cluster\_memb4,cluster\_memb4)

ari4=adj.rand.index(predicted\_cluster\_memb4,cluster\_memb4)

ri4=rand.index(predicted\_cluster\_memb4,cluster\_memb4)

**#TESTING 5 Clusters**

predicted\_cluster\_memb5<-cl\_predict(c5,test)

cluster\_memb5<-c5$cluster

table(cluster\_memb5)

predicted\_cluster\_memb5

tab5=table(predicted\_cluster\_memb5,cluster\_memb5)

ari5=adj.rand.index(predicted\_cluster\_memb5,cluster\_memb5)

ri5=rand.index(predicted\_cluster\_memb5,cluster\_memb5)

**#Creating Rand and Adjusted Rand tables**

Adjusted\_Rand\_Indexes=c(ari2,ari3,ari4,ari5)

aridf=as.data.frame(Adjusted\_Rand\_Indexes)

IDS=c('ari2','ari3','ari4','ari5')

aridf=cbind(IDS,aridf)

kable(aridf)

Rand\_indexs=c(ri2,ri3,ri4,ri5)

ris=as.data.frame(Rand\_indexs)

IDS\_r=c('RI2','RI3','RI4','RI5')

ris=cbind(IDS\_r,ris)

kable(ris)

**#basic visulisation**

ggplot(data=training,aes(y=training$GROWTH\_RATE,x=training$ALLVEHS\_AADT))+

 geom\_point(color=training$twoclusters)+

 geom\_text\_repel(,label=training$Street,color=c2m)+

 ylim(c(-3,3))+

 xlim(c(-4,4))+

 xlab("Flow of all vehicles")+

 ylab("Growth rate")

autoplot(pcaout)+

geom\_point(color=c2m)+

 geom\_text\_repel(label=training$Street)

ggplot(data=training,aes(x=training$pc1,y=training$pc2))+

geom\_point(color=factor(training$twoclusters))+

 scale\_color\_manual(values=c("blue","green"))+

 geom\_text\_repel(label=training$Street)+

 xlab("Principal Component 1")+

 ylab("Principal Component 2")+

 ggtitle("Principal Component Visulisation")

**#Dendrogram for reduced dataset**

#Reduced dataset to 20 observations for easier visualisation

Transport\_Final20=Transport\_Final[1:20,]

#create clusters

Transport\_Final20%>%

 select\_if(is.numeric)%>%

 scale%>%

 dist%>%

 hclust(method = "complete")->

 Transportcluster

#plot dendrogram

plot(Transportcluster, labels = Transport\_Final20$Streets,cex=0.7)

**#Dendrogram for roads with highest growth**

#filter data for 20 roads with highest growth.

TopG<-filter(Transport\_volume\_for\_Thea,GROWTH\_RATE>=.116)

#create clusters

TopG%>%

 select\_if(is.numeric)%>%

 scale%>%

 dist%>%

 hclust(method = "complete")->

 TopclusterG

#plot dendrogram

plot(TopclusterG,labels = TopG$Streets,cex=0.7)

**#Dendrogram for roads with lowest growth**

#filter data for 20 roads with lowest growth.

LowG<-filter(Transport\_volume\_for\_Thea,GROWTH\_RATE<(-.077))

#create clusters

LowG%>%

 select\_if(is.numeric)%>%

 scale%>%

 dist%>%

 hclust(method = "complete")->

 LowclusterG

#plot dendrogram

plot(LowclusterG,labels = LowG$Streets,cex=0.7)

**#Dendrogram for busiest roads**

#filter data for 20 busiest roads

Top<-filter(Transport\_Final,ALLVEHS\_AADT>87000)

#create clusters

Top%>%

 select\_if(is.numeric)%>%

 scale%>%

 dist%>%

 hclust(method = "complete")->

 Topcluster

#plot dendrogram

plot(Topcluster,labels = Top$Streets,cex=0.7)

**#Dendrogram for quietest roads**

#filter data for 20 quietest roads

Low<-filter(Transport\_Final,ALLVEHS\_AADT<29)

#create clusters

Low%>%

 select\_if(is.numeric)%>%

 scale%>%

 dist%>%

 hclust(method = "complete")->

 Lowcluster

#plot dendrogram

plot(Lowcluster,labels = Low$Streets,cex=0.7)